

# Numbering Positions in SIV Relative to SIVMM239 (revised\*)

Charles Calef<sup>1</sup>, John Mokili<sup>1</sup>, David H. O'Connor<sup>2</sup>, David I. Watkins<sup>2</sup>, Bette Korber<sup>1</sup>

<sup>1</sup>Theoretical Biology and Biophysics, T10, MS K710, Los Alamos National Laboratory, Los Alamos NM

87545, USA

<sup>2</sup>Wisconsin Regional Primate Research Center, 1220 Capitol Court, Madison, WI USA 53715

\* This article has been revised (Oct. 1, 2002) based on Henderson *et al.*, 1988 *J. Virol.* **62**:2587-2595.

The cleavage sites within the p2, p7, p6 and p1 segments of gag have been corrected. We thank Dr Robert J. Gorelick (Retroviral Mutagenesis Laboratory AIDS Vaccine Program) for bringing the errors to our attention.

## Introduction

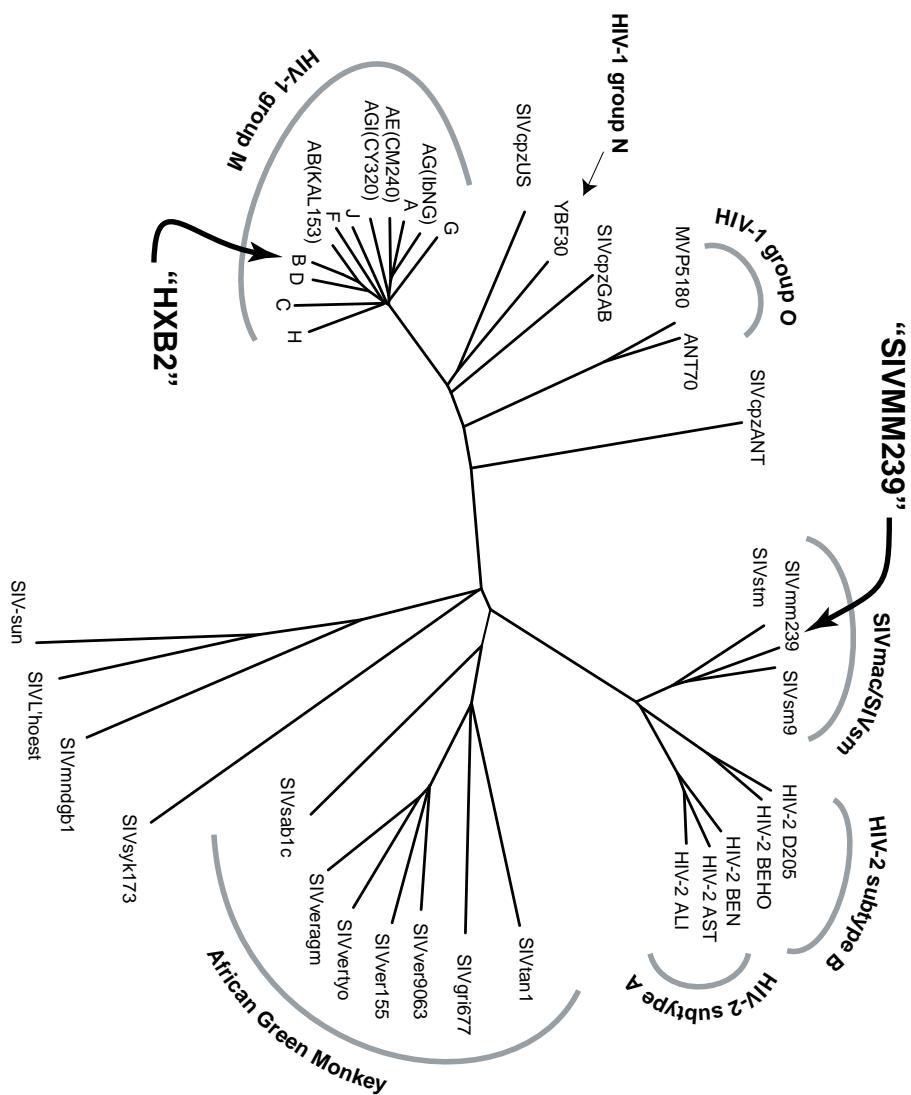
The use of HIV/HXB2 as the prototype reference strain for numbering nucleic acid and amino acid sequences has provided a useful strategy for consistent and accurate determination of the locations of nucleic and amino acid sequences of HIV-1 in the literature [1]. Because of the high frequency of insertions and deletions, different HIV sequences have genes and proteins of varying lengths. Specifying the sequence position relative to a unique reference strain, HIV/HXB2, allows direct comparisons between studies that use different strains, and easy retrieval of sequences of the gene of protein regions of interest from the databases. The HXB2 numbering engine at the Los Alamos HIV sequence database website (<http://hiv-web.lanl.gov/NUM-HXB2/HXB2.MAIN.html>) ensures that the numbering is accurate. It is very rapid and enables readers to reconstruct and reproduce what was done in an initial manuscript where HXB2 numbering was included. Specification of sequence positions is often included in papers where epitopes are defined, where primers are used, or where key functional elements are localized, and in these settings the HXB2 numbering engine is a quick way to determine the precise location of the region of interest.

This exercise is manageable for sequences that are relatively closely related to HIV/HXB2, but the more divergent the sequence under study is from HIV/HXB2, the harder it is to do the alignment to determine accurately the relative positions vis-a-vis the prototype or reference strain. HXB2 can be used readily for numbering sequences within the M group of HIV-1 viruses, and reasonably efficiently for the more diverse viral sequences from chimpanzee, and the human O and N groups (Figure 1). But the numbering of SIVs isolated from sooty mangabeys illustrates a situation where an alternative approach for numbering the nucleic and amino acid sequences is required. The deduced amino acid sequence of SIVmm239 is similar to that of SIVsmH4 by 91% in Gag, 92% in Pol, 84% in Env, 83% in Vif, 65% in Tat, 73% in Rev and 66% in Nef. Within the same regions, SIVmm239 has a similarity score of 52%, 56%, 31%, 25%, 23% 28% and 29%, respectively, to HXB2 [2]. In addition, most SIVmm, SIV and HIV-2 strains have a vpx ORF instead of vpu, a region of potential problems for numbering SIVs relative to HXB2 (Figure 2). Thus it is more practical to align and number SIVmm and HIV-2 isolates relative to a strain that has the same genomic organization and which is more closely related. Another rationale for adopting a new numbering prototype sequence for SIV is its increasing use in primate vaccine research.

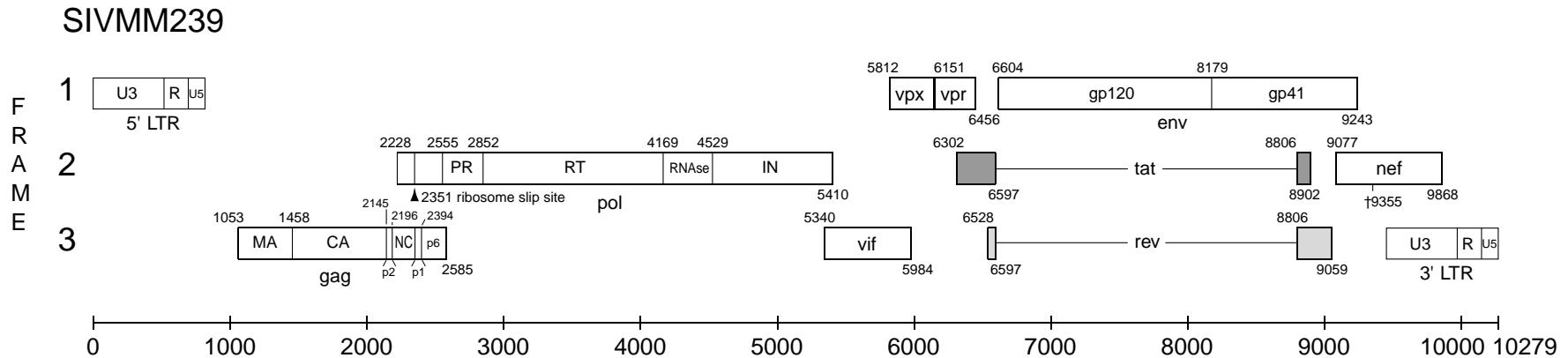
After some deliberation and external consultation, we selected SIVMM239 as the prototype reference sequence for numbering SIV strains at the Los Alamos database. There are reasonable arguments for the use of different strain as the prototype. But the high frequency with which SIVMM239 is used in vaccine studies and the comparatively large number of epitopes that have been defined for SIVMM239 was the determining factor for this choice. However, the original SIVMM239 clone [2] deposited in GenBank (accession number M33262) has 255 nucleotides of flanking non-SIVMM sequence. We have removed the flanking sequence and stored the resulting file as SIVMM239R in our database. The original sequence of SIVMM239 contains a premature stop codon, TAA, at position 9353–9355 within the nef coding sequence. In SIVMM239R we have replaced the TAA stop with the SIVMM consensus codon GAA which codes for glutamate.

In dealing with deletions and insertions relative to SIVMM239, we have used the same methodology as for the numbering of HIV-1 relative to HIVHXB2 [1]. The HXB2 numbering engine on the HIV database web site has been extended to allow SIVMM239 numbering. These tools can also be used to find positions in any reference sequence provided by the user, and as a map for finding the start and stop points of genes and proteins.

- [1] Korber, B. T., Foley, B. F., Kuiken, C. I., Pillai, S. K., and Sodroski, J. G., Numbering Positions in HIV Relative to HXB2CG, in Korber *et al.*, eds., *Human Retroviruses and AIDS 1998*, pp. III-102–III-111, Los Alamos National Laboratory, Los Alamos, NM, report LA-UR 99-1704. Available online at <http://hiv-web.lanl.gov/NUM-HXB2/NUMBERING.html>.
- [2] Regier, D. A., and Desrosiers, R. C., The Complete Nucleotide Sequence of a Pathogenic Molecular Clone of Simian Immunodeficiency Virus, *AIDS Research and Human Retroviruses*, **6**(11):1221–1231.



**Figure 1.** Phylogenetic tree of the primate lentiviruses showing the large distance between the SIVmac group and the HIV-1 M group. Note also the wide divergence of SIVmac from other SIVs.



**Figure 2. Landmarks of SIVMAC239 genome.** The gene start, indicated by the small number in the upper left corner of each rectangle normally records the position of the *a* in the *atg* start codon for that gene while the number in the lower right records the last position of the stop codon. For *pol*, the 5' end at position 2228 is the start of the open reading frame. The start of the *Pol* polyprotein is taken to be the first *t* in the sequence *tttttag* which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting -1 frameshift and the translation of the *gag-pol* polyprotein. The *tat* and *rev* spliced exons are shown as shaded rectangles. †9355 marks a premature stop codon in *nef* found in the original SIVMM239 strain sequenced and deposited in GenBank. This TAA stop codon has been replaced by a GAA glutamate codon in the reference SIVMM239 sequence annotated on the pages that follow. The putative boundaries of the constituent proteins of the *gag*, *pol*, and *env* polyproteins are tentative having been selected partly by alignment with HIV-1 strain HXB2R. Abbreviations: *MA* matrix, *CA* capsid, *NC* nucleocapsid, *PR* protease, *RT* reverse transcriptase, *IN* integrase.

## SMM239 Amino Acid Sequence Numbering:

### Gag Pr55 Gag precursor (Assemblin)

MGVRNSVLSG KKADELEKIR LRPNGKKYM LKHVVWAANE LDRFGLAESL LENKEGCQKI LSVLAPLVPT GSENLKSLYN TVCVIWCIAH EEKVKHTEEA 100  
 KQIVQRHLVV ETGTTETMPK TSRPTAPSSG RGGNYPVQQI GGNYVHLPLS PRTLNAAWKL IEEKKFGAEV VPGFQALSEG CTPYDINQML NCVGDHQAM 200  
 QIIRDIINNE AADWDLQHPQ PAPQQGQLRE PSGSDIAGTT SSVDEQIQWM YRQQNPIPVG NIYRRWIQLG LQKCVRMYNP TNILDVKQGP KEPFQSYYVDR 300  
 FYKSLRAEQT DAAVKNWMTQ TLLIQNANPD CKLVLKGGLGV NPTLEEMLT A CGVGPGQK ARLMAEALKE ALAPVPIPFA AAQQRGPRKP IKCWNCGKEG 400  
 HSARQCRAAPR RQGCWKCGKM DHVMAKCPDR QAGFLGLGPW GKKPRRNFPMA QVHQGLMPTA PPEDPAVDLL KNYMQLGKQQ REKQRESREK PYKEVTEDLL 500  
 HLNSLFGGDQ 510

### Gag p17 Matrix

MGVRNSVLSG KKADELEKIR LRPNGKKYM LKHVVWAANE LDRFGLAESL LENKEGCQKI LSVLAPLVPT GSENLKSLYN TVCVIWCIAH EEKVKHTEEA 100  
 KQIVQRHLVV ETGTTETMPK TSRPTAPSSG RGGNY 135

### Gag p24 Capsid

PVQQIGGNYV HLPLSPRTLW AWVKLIEEKK FGAEVVPGFQ ALSEGCTPYD INQMLNCVGD HQAAMQIIRD IINEEAADWD LQHPQPAPQQ GQLREPSGSD 100  
 IAGTTSSVDE QIQWMYRQQN PIPVGNIYRR WIQLGLQKCV RMYNPTNILD VKQGPKEPFQ SYVDRFYKSL RAEQTDAAVK NWMTQTLIQL NANPDCKLVL 200  
 KGLGVNPTLE EMLTACQGVG GPGQKARLM 229

### Gag p2 "Spacer"

AEALKEALAP VPIPFAA 17

### Gag p7 Nucleocapsid (NC)

AQQRGPRKPI KCWNCGKEGH SARQCRAAPR QGCWKCGKMD HVMAKCPDRQ AG 52

### Gag p1 "Spacer"

FLGLGPWGKK PRNF 14

### Gag p6

PMAQVHQGLM PTAPPEDPAV DLLKNYMQLG KQQREKQRES REKPYKEVTE DLLHLNSLFG GDQ 63

### Pol polyprotein

FFRPWSMGKE APQFPHGSSA SGADANCSPR GPSCGSAKEL HAVGQAAERK AERKQREALQ GGDRGFAAPQ FSLWRRPVVT AHIEQOPVEV LLDTGADDI 100  
 VTGIELGPHY TPKIVGGIGG FINTKEYKNV EIEVLGKRIK GTIMTGDTPN NIFGRNLLTA LGMSLNFPNA KVEPVKVALK PGKDGPKLQ WPLSKEKIVA 200  
 LREICEKMEK DGQLEEAPPT NPYNTPTFAI KKKDKNKWRM LIDFRELNRV TQDFTEVQLG IPHPAGLAKR KRITVLDIGD AYFSIPLDEE FRQYTAFTLP 300  
 SVNNAEPGKR YIYKVLPQGW KGSPAIFQYT MRHVLEPFRK ANPDVTLVQY MDDILIASDR TDLEHDRVVL QSKELLNSIG FSTPEEKFQK DPPFQWMGYE 400

LWPTKWKLQK IELPQRETWT VNDIQKLVGV LNWAQIYPG IKTKHLCLI RGKMTLTEEV QWTEMAEAAY EENKIILSQE QEGCYYQEGK PLEATVIKSQ 500  
 DNQWSYKIHQ EDKILKVGKF AKIKNTHTNG VRLLAHVIQK IGKEAIVIWG QVPKFHLPVE KDVWEQWWTD YWQVTWIPEW DFISTPPLVR LVFNLVKDPI 600  
 EGEETYYTDG SCNKQSKEGK AGYITDRGKD KVVKVLEQTNN QQAELEAFM ALTDSPKAN IIVDSQYVMG IITGCPTES SRLVNQIIIE MIKKSEIYVA 700  
 WVPAHKGIGG NQEIDHLVSQ GIRQVLFLEK IEPAQEEHDK YHSNVKELVF KFGLPRIVAR QIVDTCDKCH QKGEAIHGQA NSDLGTWQMD CTHLEGKIII 800  
 VAVHVASFV EAEVIPQETG RQTAFLLLKL AGRWPITHLH TDNGANFASQ EVKVMVAWWAG IEHTFGVPYN PQSQGVVEAM NHHLKNQIDR IREQANSVET 900  
 IVLMAVHCMN FKRRGGIGDM TPAERLINMI TTEQEIQFQQ SKNSKFKNFR VYYREGRDQL WKGPGEELLWK GEGAVILKVG TDIKVVPRRK AKIICKDYGGG 1000  
 KEVDSSSHME DTGEAREVA 1019

**Pol p10 Protease**

PQFSLWRRPV VTAHIEGQPV EVLLDTGADD SIVTGIELGP HYTPKIVGGI GGFINTKEYK NVEIEVLGKR IKGTIMTGDT PINIFGRNLL TALGMSLNF 99

**Pol p66 Reverse Transcriptase (RT/RNAse)**

PIAKVEPVKV ALKPGKDGPK LKQWPLSKEK IVALREICEK MEKDGQLEEA PPTNPYNTPT FAIKKKDKNK WRMLIDFREL NRVTQDFTEV QLGIPHPAGL 100  
 AKRKIRTVLD IGDAYFSIPL DEEFRQYTAF TLPSVNNAEP GKRYIYKVL P QGWKGSPAIF QYTMRHVLEP FRKANPDVTL VQYMDILIA SDRTDLEHDR 200  
 VVLQSKELLN SIGFSTPEEK FQKDPPFQWM GYELWPTKW LQKIELPQRE TWTVNDIQKL VGVLNWAAQI YPGIKTKHLC RLIRGKMLT EEVQWTEMAE 300  
 AEYEENKIIL SQEQEGCYYQ EGKPLEATVI KSQDNQWSYK IHQEDKILKV GKFAKIKNTH TNGVRLLAHV IQKIGKEAIV IWGQVPKFHL PVEKDVWEQW 400  
 WTDYWQVTWI PEWDFISTPP LVRLVFNVLV DPIEGETYY TDGSCNKQSK EGKAGYITDR GKDKVKVLEQ TTNNQQAELA FLMALTDSPG KANIIVDSQY 500  
 VMGIITGCPT ESESRLVNQI IEEMIKKSEI YVAWVPAHKG IGGNQEIDHL VSQGIRQVL 559

**Pol p51 RT**

PIAKVEPVKV ALKPGKDGPK LKQWPLSKEK IVALREICEK MEKDGQLEEA PPTNPYNTPT FAIKKKDKNK WRMLIDFREL NRVTQDFTEV QLGIPHPAGL 100  
 AKRKIRTVLD IGDAYFSIPL DEEFRQYTAF TLPSVNNAEP GKRYIYKVL P QGWKGSPAIF QYTMRHVLEP FRKANPDVTL VQYMDILIA SDRTDLEHDR 200  
 VVLQSKELLN SIGFSTPEEK FQKDPPFQWM GYELWPTKW LQKIELPQRE TWTVNDIQKL VGVLNWAAQI YPGIKTKHLC RLIRGKMLT EEVQWTEMAE 300  
 AEYEENKIIL SQEQEGCYYQ EGKPLEATVI KSQDNQWSYK IHQEDKILKV GKFAKIKNTH TNGVRLLAHV IQKIGKEAIV IWGQVPKFHL PVEKDVWEQW 400  
 WTDYWQVTWI PEWDFISTPP LVRLVFNVLV DPIEGETY 439

**Pol p15 RNase**

YTDGSCNKQS KEGKAGYITD RGKDKVKVLE QTTNQQAELA AFLMALTDSPG PKANIIIVDSQ YVMGIITGCP TESESRLVNQ IIEMIKKSE IYVAWVPAHK 100  
 GIGGNQEIDH LVSQGIRQVL 120

**Pol p31 Integrase**

FLEKIEPAQE EHDKYHSNVK ELVFKFGLPR IVARQIVDTC DKCHQKGEAI HQQANSILGT WQMDCTHLEG KIIIVAVHVA SGFIEAEVIP QETGRQTAFL 100  
 LLKLAGRWPI THLHTDNGAN FASQEVKMVA WWAGIEHTFG VPYNPQSQGV VEAMNHHLKN QIDRIREQAN SVETIVLMAV HCMNFKRRGG IGDMTPAERL 200  
 INMITTEQEI QFQQSKNSKF KNFRVYYREG RDQLWKGPGF LLWKGEAVI LKVGTDIKVV PRRKAKIICD YGGGKEVDSS SHMEDTGEAR EVA 293

**Vif**

MEEEKRWIAV PTWRIPERLE RWHSLIKYLK YKTKDLQKVC YVPHFKVGWA WWTCSRVIFF LQEGLSHLEVQ GYWHLTPEKG WLSTYAVRIT WYSKNFWTDV 100  
 TPNYADILH STYFPCFTAG EVRRAIRGEQ LLSCCRFPRA HKYQVPSLQY LALKVVSDVR SQGENPTWKQ WRRDNRRGLR MAKQNSRGDK QRGGKPPTKG 200  
 ANFPGLAKVL GILA 214

**Vpx**

MSDPRERIIPP GNSGEETIGE AFEWLNRTE EINREAVNHL PRELIFQVWQ RSWEYWHDEQ GMSPSYVKYR YLCLIQKALF MHCKKGCRCL GEGHGAGGWR 100  
 PGPPPPPPPG LA 112

**Vpr**

MEERPPNEG PQREPWDEWV VEVLEELKEE ALKHFDPRLL TALGNHIYNR HGDTLEGAGE LIRILQRALF MHFRGGCIHS RIGQPGGGNP LSAIPPSRSM 100  
 L 101

**Tat**

METPLREQEN SLESSNERSS CISEADASTP ESANLGEEL SQLYRPLEAC YNTCYCKKCC YHCQFCFLKK GLGICYEQSR KRRRTPKKAK ANTSSASNKP 100  
 ISNRTRHCQP EKAKKETVEK AVATAPGLGR 130

**Rev**

MSNHEREEEEL RKRLRLIHLL HQTNPYPTGP GTANQRRQRK RRWRRRWQQL LALADRIYSF PDPPTDTPLD LAIQLQNLIAIESIPDPPTN TPEALCDPTE 100  
 DSRSPQD 115

**Env**

MGCLGNQLLI AILLLSVYGI YCTLYVTVFY GVPAWRNATI PLFCATKNRD TWGTTQCLPD NGDYSEVALN VTESFDAWNN TVTEQAIEDV WQLFETSIKP 100  
 CVKLSPLCIT MRCNKSETDR WGLTKSITT ASTTSTTASA KVDMVNETSS CIAQDNCTGL EQEQMISCKF NMTGLKRDKK KEYNETWYSA DLVCEQGNNT 200  
 GNESRCYMH CNTSVIQESC DKHYWDAIRF RYCAPPGYAL LRCNDTNYSG FMPKCSKVVV SSCTRMMETQ TSTWFGFNGT RAENRTYIYW HGRDNRTIIS 300  
 LNKYYNLTMK CRRPGNKTBL PVTIMSGLVF HSQPINDRPK QAWCWFGGW KDAIKEVKQT IVKHPRTGT NNTDKINLTA PGGGDPETVF MWTNCRGEFL 400  
 YCKMNWFLNW VEDRNTANQK PKEQHKRNYY PCHIRQIINT WHKVGKNVYL PPREGDLTCN STVTSЛИANI DWIDGNQTNI TMSAEVAELY RLELDYKLV 500

gp120 end \ gp41 start

EITPIGLAPT DVKRYTTGGT SRNKRGVVFVLFGLGFLATAG SAMGAASLTL TAQSRTLLAG IVQQQQQLLD VVKRQQELLR LTVWGKKNLQ TRVTAIEKYL 600  
 KDQAQLNAWG CAFRQVCHTT VPWPNASLTP KWNNETWQEW ERKVDFLEEN ITALLEEAQI QQEKNMYLEQ KLNSWDVFGN WFDLASWIKY IQYGVYIVVG 700  
 VILLRIVIYI VQMLAKLRQG YRPVFSSPPS YFQQTHIQQD PALPTREGKE RDGGEGGGNS SWPWQIEYIH FLIRQLIRLL TWLFSNCRTL LSRYQILQP 800  
 ILQRLSATLQ RIREVLRTEL TYLQYGWSYF HEAVQAVWRS ATETLAGAWG DLWETLRRGG RWILAIPRRI RQGLELTL 879

Premature stop in original SIVMM239 sequence,  
 changed to consensus glutamate, E.

**Nef**

MGGAIMSRRS RPSGDLRQRL LRARGETYGR LLGEVEDGYS QSPGGLDKGL SSLSCEGQKY NQGQYMNTPW RNPAEEREKL AYRKQNMDDI DEEDDDLVGV 100  
 SVRPKVPLRT MSYKLAIDMS HFIKEKGGL GIYYSARRHR ILDIYLEKEE GIIPDWQDYT SGPGIRYPKT FGWLWKLVPV NVSDEAQEDE EHYLMHPAQT 200  
 SQWDDPWGEV LAWKFDPTLA YTAEAYVRYP EEFGSKSGLS EEEVRRRLTA RGLLNMADKK ETR 263

## SMM239 Nucleic Acid Sequence Numbering:

/ 5' LTR U3 region start

tggaaaggat ttattacagt gcaagaagac atagaatctt agacatatac ttagaaaagg aagaaggcat cataccagat tggcaggatt acacctcagg 100  
 accaggaatt agataccaa agacatttg ctggctatgg aaattagtcc ctgtaaatgt atcagatgag gcacaggagg atgaggagca ttatthaatg 200  
 catccagctc aaactccca gtggatgac cttggggag aggttctagc atgaaagtt gatccaactc tggcctacac ttatgaggca tatgttagat 300  
 acccagaaga gtttggaaagc aagtcaaggcc tgtcagagga agaggttaga agaaggctaa ccgcaagagg ccttcttaac atggctgaca agaaggaaac 400  
 tcgctgaaac agcaggact ttccacaagg ggtatgtacg gggaggtact ggggaggagc cggtcgggaa cgcccactt cttgtatgtat aaatatcact 500

5' LTR U3 region end \ / 5' LTR R repeat region start

/ putative mRNA start

gcatttcgct ctgtattcag tcgctctcg gaggagctgg cagattgagc cctggaggt tctctccagc actagcaggt agagcctggg tggccctgc 600  
 5' LTR U5  
 5' LTR R repeat region end \ \ region start  
 tagactctca ccagcacttg gccgggtctg ggcagagtga ctccacgctt gcttgcttaa agccctcttc aataaagctg ccattttaga agtaagctag 700  
 tgggtgttcc catctctccct agccggccgc tggtaactc ggtactcaat aataagaaga ccctggcttg tttaggaccct ttctgccttg ggaaaccgaa 800  
 gcaggaaaaat cccttagcaga ttggggcctg aacaggact tgaaggagag tgagagactc ctgagtaacgg ctgagtgaag gcagtaaggg cggcaggaaac 900  
 caaccacgac ggagtgcctcc tataaaggcg cgggtcggtt ccagacggcg tgaggagcgg gagaggaaga ggcctccggg tgcaggtaa tgcaacaccaa 1000

/ Gag p17 start

aaaagaaaata gctgtctttt atccaggaag gggtaataag atagagtggg agatgggctg gagaactcc gtcttgtcag ggaagaaaagc agatgaatta 1100  
 gaaaaaatta ggctacgacc caacggaaag aaaaagtaca tggtaagca tggtagtatgg gcagcaaatg aatttagatag atttggatta gcagaaagcc 1200  
 tggggagaa caaagaagga tggtaaaaaa tactttcggt cttagctcca tttagtgc当地 caggtcaga aaattttata atactgtctg 1300  
 cgtcatctgg tgcatttcacg cagaagagaa agtgaacac actgaggaag caaaacagat agtgcagaga cacctgtgg tggaaacacagg aacaacagaa 1400

Gag p17 end \ / Gag p24 start

actatgccaa aaacaagttag accaacagca ccatctagcg gcagaggagg aaattaccctt gtacaacaaa taggtggtaa ctatgtccac ctgccattaa 1500  
 gcccggaaac attaaatgcc tgggtaaaat tggtagagga aaagaaattt ggagcagaag tagtgc当地 ctttcaggca ctgtcagaag gttgcacccc 1600  
 ctatgacatt aatcagatgt taaaattgtgt gggagaccat caagcggcta tggtagattt cagagatatt ataaacgagg aggctcaga ttgggacttg 1700  
 cagcacccac aaccagctcc acaacaagga caacttaggg agccgtcagg atcagatatt gcaggaacaa ctagttcagt agatgaacaa atccagtgg 1800  
 tggtagacaca acagaacccc ataccatgt gcaacatttt cagggatgg atccaaactgg gtttgc当地 atgtgtcaga atgtataacc caacaaacat 1900  
 tcttagatgtt aaacaaggcc caaaagagcc atttcagagc tatgtagaca ggttctacaa aagtttaaga gcagaacaga cagatgc当地 agttaagaat 2000  
 tggatgactc aaacactgct gattcaaaat gctaaccag attgcaagct agtgc当地 gggctgggtt tgaatcccac cctagaagaa atgtgc当地 2100

Gag p24 end \ / Gag p2 start

Gag p2 end \ / Gag NC (p7) start

cttgc当地 agtagggggg cccggacaga aggtagatt aatggcagaa gccc当地 gcca accagtgc当地 atcccttttgc当地 cagc当地 cccca 2200  
 acagagggga ccaagaaagc caatgttgc ttggattgt gggaaagagg gacactctgc aaggcaatgc agagccccaa gaagacaggg atgtggaaa 2300

Gag NC (p7) end \/\ Gag p1 start

ribosome -1 slip site Gag to Gag-Pol

/ Pol start

Gag p1 end \/\ Gag p6 start

tgtggaaaaa tggaccatgt tatggccaaa tgcccagaca gacaggcggg tttttttagc cttggtccat ggggaaagaa gcccccgaat ttccccatgg 2400  
ctcaagtgc a tcaggggctg atgccaactg ctccccaga ggacccagct gtggatctgc taaagaacta catgcagttg ggcaaggcgc agagagaaaa 2500

/ Pol protease start

Gag p6 end \

gcagagagaa agcagagaga agccttacaa ggaggtgaca gaggattgc tgcaccta ttctctctt ggaggagacc agtagtcact gctcatattg 2600  
aaggacagcc tgtagaagta ttactggata caggggctga tgattctatt gtaacagggaa tagagtttagg tccacattat accccaaaaa tagtaggagg 2700  
aataggaggt ttatataata ctaaagaata caaaaatgt a gaaatagaag ttttaggcaa aaggattaaa gggacaatca tgacagggga caccggatt 2800

Pol protease end \/\ Pol p66 & p51 RT start

aacattttg gtagaaattt gctaacagct ctggggatgt ctctaaattt tcccatagct aaagtagagc ctgtaaaagt cgccctaaag ccaggaaagg 2900  
atggacaaa attgaagcag tggccattat caaaaagaaaa gatagttgca ttaagagaaaa tctgtaaaaa gatggaaaag gatggctagt tggaggaagc 3000  
tccccgacc aatccatatac acaccccccac atttgcata aagaaaaagg ataagaacaa atggagaatg ctgatagatt ttagggact aaatagggtc 3100  
actcaggact ttacggaagt ccaatttagga ataccacacc ctgcaggact agcaaaaagg aaaagaatta cagtaactgga tataggtgat gcatatttct 3200  
ccatacctt agatgaagaa tttaggcagt acactgcctt tacatttacca tca gtaata atgcagagcc aggaaaacga tacatttata aggttctgcc 3300  
tcagggatgg aagggtcac cagccatctt ccaatacact atgagacatg tgctagaacc ctccaggaag gcaaatccag atgtgacctt agtccagtagt 3400  
atggatgaca tcttaatagc tagtgacagg acagacttgg aacatgacag ggtagtttta cagtc当地agg aactcttgcg tagcataggg ttttctaccc 3500  
cagaagagaa attccaaaaa gatccccat ttcaatggat ggggtacgaa ttgtggccaa caaaaatggaa gttgcaaaag atagaggc cacaagaga 3600  
gacctggaca gtgaatgata tacagaagtt agtaggagta ttaaatttggg cagctcaaat ttatccaggt ataaaaacc aacatctcg taggttaatt 3700  
agaggaaaaa tgactcta ac agaggaagtt cagtgactg agatggcaga agcagaatata gaggaaaata aaataattct cagtc当地ggaa caagaaggat 3800  
gttattacca agaaggcaag ccatttagaag ccacggtaat aaagagtcg gacaatcagtt ggtcttataa aattcaccaaa gaagacaaa tactgaaagt 3900  
aggaaaaattt gcaaaagataa agaatacaca taccatgg a gtgagactat tagcacatgt aatacagaaa atagggaaagg aagcaatagt gatctggga 4000  
caggtcccaa aattccactt accagtttag aaggatgtat gggAACAGTG gtggacagac tattggcagg taacctggat accggaaatgg gattttatct 4100

Pol p51 end p66 RT continues \/\ Pol p15 RNase start

caacaccacc gctagtaaga ttagtctca atctagtgaa ggaccctata gagggagaag aaaccttata tacagatgg tcatgtaata aacagtcaaa 4200  
agaaggaaaa gcaggatata tcacagatag gggcaaagac aaagttttttt tgtagaaaca gactactaat caacaaggcag aattggaaagc atttctcatg 4300  
gcattgacag actcaggccc aaaggcaat attatagtag attcacaata tggtatggg ataataacag gatgccctac agaatcagag agcaggctag 4400  
ttaatcaaat aatagaagaa atgataaaaa agtcagaaat ttatgtgca tgggtaccag cacacaaagg tataggagga aaccaagaaa tagaccaccc 4500

Pol p15 RNase, p66 RT end \/\ Pol p31 integrase start

agtttagtcaa gggatttagac aagttcttctt cttggaaaaag atagagccag cacaagaaga acatgataaa taccatgta atgtaaaaga attggattc 4600  
aaatttggat taccatggat agtggccaga cagatagtag acacctgtga taaatgtat cagaaaggag aggctataca tgggcaggca aattcagatc 4700  
tagggacttg gcaaatggat tgtagccatc tagagggaaa aataatcata gttgcagtag atgttagctg tggattcata gaagcagagg taattccaca 4800  
agagacagga agacagacag cactattct gttaaaatttgc gcaaggcagat ggcctattac acatctacac acagataatg gtgctaactt tgcttcgca 4900  
gaagtaaaga tgggtgcata gttggcaggat atagagcaca cctttgggtt accatacaat ccacagatc agggagtagt ggaagcaatg aatcaccacc 5000

tgaaaaatca aatagataga atcagggaaac aagcaaattc agtagaaacc atagtattaa tggcagttca ttgcataat tttaaaagaa ggggaggaat 5100  
 aggggatatg actccagcag aaagattaat taacatgtc actacagaac aagagataca atttcaacaa tcaaaaaact caaaatttaa aaatttcgg 5200  
 gtctattaca gagaaggcag agatcaactg tggagggac ccgtgagct attgtggaaa ggggaaggag cagtcataatc aaagtaggg acagacatta 5300

## / Vif start

aggtagtacc cagaagaaag gctaaaatta tcaaagatta tggaggagga aaagaggtgg atagcagttc ccacatggag gataccggag aggctagaga 5400

Pol, Gag-Pol, and  
p31 integrase end \

ggtggcatag cctataaaaa tatctgaaat ataaaactaa agatctacaa aaggttgt atgtccccca ttttaaggc ggtggccat ggtggacctg 5500  
 cagcagagta atcttcccac tacaggaagg aagccattta gaagtacaag ggtattggca tttgacacca gaaaaagggt ggctcgtac ttatgcgtg 5600  
 aggataacct ggtactcaa gaactttgg acagatgtaa cacaaacta tgcagacatt ttactgcata gcacttattt cccttgcattt acagcgggag 5700  
 aagtgagaag ggccatcagg ggagaacaac tgctgtttt ctgcaggttc ccgagagctc ataagtagcca ggtaccaagc ctacagtact tagcactgaa 5800

## / Vpx start

agtagtaagc gatgtcagat cccagggaga gaatcccacc tggaaacagt ggagaagaga caataggaga ggccttcgaa tggctaaaca gaacagtaga 5900

## Vif end \

ggagataaac agagaggcgg taaaccaccc accaaggagg ctaatttcc aggtttggca aaggcttgg gaatactggc atgatgaaca agggatgtca 6000  
 ccaagctatg taaaatacag atacttgtgt ttaatacataa aggcttattt tatgcattgc aagaaaggct gtagatgtct aggggaagga catggggcag 6100

## Vpx end \ / Vpr start

ggggatggag accaggaccc ctcctccctc cccctccagg actagcataa atgaaagaaa gacctccaga aatgaaggg ccacaaagg aaccatggg 6200  
 tgaatggta gtggaggttc tggagaact gaaagaagaa gctttaaac attttgcattcc tcgcttgcta actgcacttg gtaatcatat ctataataga 6300

## / Tat exon 1 start

catggagaca cccttgagg agcaggagaa ctcattagaa tcctccaacg agcgctctc atgcattca gaggcggatg catccactcc agaatcggcc 6400

## Vpr end \

aacctgggg aggaaatcct ctctcagcta taccgcctc tagaagcatg ctataacaca tgctattgtt aaaagtgtt ctaccattgc cagttttgtt 6500

/ Rev exon 1 start Tat, Rev exon 1 end \ / Tat, Rev intron  
 ttcttaaaaa aggcttgggg atatgttatg agcaatcagc aaagagaaga agaactccga aaaaggctaa ggctaataca tcttctgcat caaacaagta 6600

## / Env gp120, gp160 start, signal peptide

agtatggat gtcttggaa tcagctgtt atcgccatct tgcttttaag tgtctatggg atctattgtt ctcttatgtt cacagtctt tatgggtac 6700  
 cagcttggag gaatgcgaca attccctct tttgtcaac caagaatagg gatacttggg gaacaactca gtgcctacca gataatggg attattcaga 6800

agtggccctt aatgttacag aaagcttga tgcctggaat aatacagtca cagaacaggc aatagaggat gtatggcaac tctttgagac ctcataaaag 6900  
 ccttgttaa aattatcccc attatgcatt actatgagat gcaataaaaag tgagacagat agatggggat tgacaaaatc aataacaaca acagcatcaa 7000  
 caacatcaac gacagcatca gcaaaaagtag acatggcaa tgagactagt tcttgatag cccaggataa ttgcacaggc ttggaacaag agcaaatgtat 7100  
 aagctgtaaa ttcaacatga cagggttaaa aagagacaag aaaaaagagt acaatgaaac ttggactct gcagatttg tatgtgaaca agggataaac 7200  
 actggtaatg aaagtagatg ttacatgaac cactgtaaaca ctctgttat ccaagagatct tgtgacaaac attattggta tgctattaga ttttagttt 7300  
 gtgcacccc aggttatgtct ttgcttagat gtaatgacac aaattattca ggctttagtgc ctaaatgttc taaggtgggt gtctctcat gcacaaggat 7400  
 gatggagaca cagacttcta ctgggttgg cttaatgga actagagcg aaaatagaac ttatattac tggcatggta gggataatag gactataatt 7500  
 agtttaataa agtattataa tctaacaatg aaatgttagaa gaccaggaaa taagacagt ttaccagtca ccattatgtc tggattggtt ttccactcac 7600  
 aaccaatcaa tgataggcca aagcaggcat ggtgttggg tggagggaaa tggaggatg caataaaaaga ggtgaagcg accattgtca aacatcccg 7700  
 gtatactgga actaacaata ctgataaaaat caatgttgcg gctctggag gaggagatcc ggaagttacc ttcatgtgga caaattgcag aggagatcc 7800  
 ctctactgta aaatgaattt gtttctaaat tggtagaag ataggaatac agctaaccag aagccaaagg aacagcataa aaggaattac gtgccatgtc 7900  
 atattagaca aataatcaac acttggcata aagtaggcaa aaatgtttat ttgcctccaa gagagggaga cctcacgtgt aactccacag tgaccagtct 8000  
 catagcaaac atagattgga ttgtatggaa ccaaactaat atcaccatga gtgcagaggt ggcagaactg tatcgattgg aattgggaga ttataaatta 8100

## Env gp120 end \ / Env gp41 start

gtagagatca ctccaaattgg ctggcccccc acagatgtga agaggtacac tactggtggc acctaagaa ataaaagagg ggtctttgt cttagggttct 8200  
 tgggttttct cgcaacggca ggttctgaa tggcgccggc gtcgttgacg ctgaccgtc agtcccgAAC tttattggct gggataatgtc agcaacacgca 8300  
 acagctgttgc gacgtggtca agagacaaca agaattgttgcg cgaactgaccg tctggggaaac aaagaacctc cagacttaggg tcactgccc catcgaaatgtac 8400  
 ttaaaggacc aggcgcagct gaatgttgcg ggtatgtcgt ttagacaagt ctgccacact actgtaccat gccaaatgtc aagtctaaaca ccaaagtgg 8500  
 acaatgagac ttggcaagag tgggagcgaa aggttgactt ctggaaagaa aatataacag ccctcctaga ggaggcccaa attcaacaag agaagaacat 8600  
 gtatgaattt caaaagttga atagctgggaa tggatgttgc aattggtttgc accttgcttcc ttggataaag tatataacaat atggagtttataatgtt 8700  
 ggagtaatac tgtaaagaat agtgcattat atagtacaaa tgctagctaa gttaaaggcag gggtagatggc cagtgttcttcc ttcccccaccc tcttatttcc 8800

## Tat, Rev

intron end \ / Tat, Rev exon 2 start

agcagaccca tatccaaacag gacccggcac tgccaaaccag agaaggcaaa gaaagagacg gtggagaagg cggtgccaaac agtcctggc cttggcagat 8900

## Tat exon 2

end \

agaatatatt catttcctga tccgccaact gatacgccctc ttgacttggc tattcagcaa ctgcagaacc ttgctatcgaa gagtatacca gatcctccaa 9000

## Rev exon 2 end \

/ Nef start

ccaaatactcc agaggctctc tgcgacccta cagaggattc gagaagtcct caggactgaa ctgaccctacc tacaatatgg gtggagctat ttccatgagg 9100  
 cggtccaggc cgtctggaga tctgcgacag agactttgc gggcgctgg ggagacttat gggagactct taggagaggt ggaagatggta tactcgaaat 9200

## Env gp41, gp160 end \

ccccaggagg attagacaag ggcttgagct cactctttg tgagggacag aaataacaatc agggacagta tatgaatact ccatggagaa acccagctga 9300

Premature in-frame stop taa in  
original SIVMM239 sequence

```

agagagagaa aaatttagcat acagaaaaaca aaatatggat gatatacatg aggaagatga tgacttggta ggggtatcag tgaggccaaa agttccccta 9400
                                                 / 3'LTR U3 region start
agaacaatga gttacaaatt ggcaatagac atgtctcatt ttataaaaga aaagggggga ctggaaaggga tttattacag tgcaagaaga catagaatct 9500
tagacatata cttagaaaag gaagaaggca tcataccaga ttggcaggat tacacctcag gaccaggaat tagataccca aagacatttg gctggctatg 9600
gaaatttagtc cctgtaaatg tatcagatga ggcacaggag gatgaggagc attatthaat gcatccagct caaactccc agtggatga cccttgggga 9700
gaggttctag catggaagtt tgcattcaact ctggcctaca cttatgaggc atatgtttaga taccagaag agtttggaaag caagtcaaggc ctgtcagagg 9800
aagaggttag aagaaggcta accgcaagag gccttcttaa catggctgac aagaaggaaa ctcgtgaaa cagcaggagc tttccacaag gggatgttac 9900

                                                 3'LTR U3 region end \ / 3'LTR R repeat start
ggggaggtac tggggaggag ccggtcggga acgcccactt tcttgatgta taaatatcac tgcatttcgc tctgtattca gtgcgtctgc ggagaggctg 10000
gcagatttag ccctggggagg ttctctccag cactagcagg taggcctgg gtgttccctg ctagactctc accagcactt ggccggtgct gggcagagtg 10100

                                                 3'LTR repeat end \ \ 3'LTR U5 region start
actccacgct tgcttgctta aagccctttt caataaaagct gccattttag aagtaagcta gtgtgtgttc ccatctctcc tagccggccgc ctggtaact 10200
cggtactcaa taataagaag accctggtct gttaggaccc tttctgcttt gggaaaccga agcaggaaaa tccctagca                                         10279

```